ELSEVIER

Computer Methods and Programs in Biomedicine 70 (2003) 187–197

Computer Methods and Programs in Biomedicine

www.elsevier.com/locate/cmpb

Assessment of the mean-value of 17-hydroxypregnenolone in the umbilical blood of newborns by the exploratory analysis of biochemical data

Milan Meloun [a,*], Martin Hill [b], Jiří Militký [c], Karel Kupka [d]

[a] Department of Analytical Chemistry, Faculty of Chemical Technology, Pardubice University, CZ-532 10 Pardubice, Czech Republic
[b] Institute of Endocrinology, Národní 8, CZ-116 94 Prague 1, Czech Republic
[c] Department of Textile Materials, Technical University, CZ-461 17 Liberec, Czech Republic
[d] Trilobyte Statistical Software Ltd., CZ-530 02 Pardubice, Czech Republic

Received 10 February 2001; received in revised form 10 November 2001; accepted 24 December 2001

Abstract

The main aim of data analysis in biochemical metrology is the extraction of relevant information from biochemical data measurements. A system of extended exploratory data analysis (EDA) based on the concept of graphical tools for sample data summarization and exploration is proposed and the original EDA algorithm in S-Plus is available on the Internet at http://www.trilobyte.cz/EDA. To check basic assumptions about biochemical and medical data is to examine the independence of sample elements, sample normality and homogeneity. The exact assessment of the mean-value and the variance of steroid levels in controls is necessary for the correct assessment of the samples from patients. Data examination procedures are illustrated by a determination of the mean-value of 17-hydroxypregnenolone in the umbilical blood of newborns. For an asymmetric, strongly skewed sample distribution corrupted with outliers the best estimate of location seems to be the median. The Box–Cox transformation improves a sample symmetry. The proposed procedure gives reliable estimates of a mean-value for an asymmetric distribution of 17-hydroxypregnenolone when the arithmetic mean can not be used. © 2002 Elsevier Science Ireland Ltd. All rights reserved.

Keywords: Exploratory data analysis; Box–Cox transformation; Power transformation; Normality test; Independence test; Homogeneity test; Robust estimates; Chemometrics; Biometrics

1. Introduction

Metrology is one of those important disciplines enabling the evaluation of measurements and observations. Traditionally, the core of metrology lies in the realization of measurements. Statistical tests and interval estimates are designed to yield

* Corresponding author. Tel.: +420-40-603-7026; fax: +420-40-603-7068.
E-mail address: milan.meloun@upce.cz (M. Meloun).

0169-2607/02/$ - see front matter © 2002 Elsevier Science Ireland Ltd. All rights reserved.
PII: S0169-2607(02)00013-5

reliable results with biological, biochemical and clinical data which meet certain requirements [1–7]. These requirements are represented by certain assumptions about the nature of the data, or observations. If our data do not meet the assumptions, the results may give incorrect answers. The most usual assumptions are: independence of the observations, normal distribution of errors and absence of gross errors (outliers). If one observation does not affect other observations, the observations are said to be independent.

When an exploratory data analysis (EDA) indicates that the sample distribution strongly differs from a symmetrical and normal one, we are faced with the problem of how to analyze the data. We transform the data by applying a single mathematical function to all of the raw data values. The reasons for transforming original data include transformation for symmetry. Symmetry in a data batch is often a desirable property, as many estimates of location give the best results and are best understood when the data come from a symmetric distribution. In perfectly symmetric data, all midsums would be equal to the median. If the data are skewed to the right, the midsums increase as they come from quantiles which reflect the asymmetric behaviour of the tails. For data skewed to the left, the midsums decrease.

This paper provides a description of some statistical procedures applied for an examination of all of the above sample assumptions for biochemical or clinical data. Data examination procedures are illustrated on a case study concerning an assessment of the mean-value of 17-hydroxypregnenolone in the umbilical blood of newborns.

2. Theoretical

Statistical treatment of experimental data supposes that the data are independent random variables coming from the same distribution, obviously normal, and that the sample size is sufficient for precise estimates of location and spread to be obtained. When some of these assumptions are not met, the data analysis is rather complicated. These assumptions must be examined by confirmatory data analysis (CDA) before interval estimation and testing.

2.1. Examination for independence of sample elements

The basic assumption of good measurement is that the individual measurements, observations in the biochemical sample batch, are independent. Interdependence of measurements can obviously be caused by

1. instability of the measurement device, for example, a shift in readings with temperature;
2. variable conditions for the measurements, which could suddenly change;
3. neglect of important factor(s) which have a major influence on measurement, for example, the sample volume, temperature, purity of chemicals, etc.; and
4. false and non-random (stratified) choice of values in a sample.

When some experimental conditions change over time, a time dependence in the observations may be indicated. When there is a sudden change in observations, a heterogeneous sample is formed. In both of these cases, a higher value of variance is found than for a homogeneous sample.

A time dependence or dependence on the order of observations can be tested for by examining the significance of the first order autocorrelation coefficient ρ_a according to

$$t_n = \frac{T_1\sqrt{(n+1)}}{\sqrt{(1-T_1)}} \tag{1a}$$

where

$$T_1 = \left(1-\frac{T}{2}\right)\sqrt{\frac{n^2-1}{n^2-4}} \tag{1b}$$

and T is the von Neumann ratio defined by

$$T = \frac{\sum_{i=1}^{n-1}(x_{i+1}-x_i)^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}. \tag{1c}$$

When the null hypothesis H_0: $\rho_a = 0$ is valid, the test criterion t_n has a Student distribution with $(n+1)$ degrees of freedom. The alternative hypothesis H_A: $\rho_a \neq 0$. When $|t_n| > t_{1-\alpha/2}(n+1)$, the null hypothesis about the independence of sample observations is rejected at the significance level α.

2.2. Examination for normality of sample distribution

Normality of a sample distribution is the basic assumption for most statistical data treatment, because many statistical tests require normality. When the type of deviation from normality of the sample is known before statistical inference, *directional tests* are used; when the type of deviation from normality is unknown, the *omnibus tests* are used.

Generally, statistical tests are less sensitive to deviations from normality than are diagnostic graphs of EDA [8]. Moreover, deviation from normality can also be caused by the presence of outliers. When the normality of a sample distribution is not proven, the data should be analyzed with great care. For testing the normality of a sample distribution, the rankit plot of the EDA [8] is one of the most useful tools, but other useful tests are available.

Test of combined sample skewness and kurtosis: the testing criterion is defined as

$$C_1 = \frac{\hat{g}_1^2(x)}{D(\hat{g}_1(x))} + \frac{(\hat{g}_2(x) - 3)^2}{D(\hat{g}_2(x))} \tag{2}$$

where $\hat{g}_1(x)$ is the sample skewness and $D(\hat{g}_1(x))$ is its variance, $\hat{g}_2(x)$ is the sample kurtosis and $D(\hat{g}_2(x))$ is its variance. For a normal distribution, the test criterion C_1 has approximately an χ^2 distribution, so that when $C_1 > \chi^2_{1-\alpha}(2)$, the null hypothesis about normality of sample distribution is rejected.

2.3. Examination for minimum sample size

The sample size n has an influence on the precision of estimates, and controls the size of confidence intervals. For very small sample sizes it may happen that hypothesis tests are affected more by the sample size n than by the variability of the data. The procedure for finding the sample size that is sufficient is as follows: From n_1 starting values, the sample variance $s_0^2(x)$ is calculated. The minimum size $n_{\min}$ of a sample taken from a normal distribution is calculated in such a way that for a given probability $(1-\alpha)$ and value of d, the confidence interval will be $\mu - d \le \bar{x} \le \mu + d$. Then $n_{\min}$ is given by

$$n_{\min} = s_0^2(x)\left[\frac{t_{1-\alpha/2}(n_1 - 1)}{d}\right]^2 \tag{3}$$

where $t_{1-\alpha/2}(n_1 - 1)$ is the quantile of the Student distribution with $(n_1 - 1)$ degrees of freedom.

2.4. Examination of sample homogeneity

Sample heterogeneity becomes evident when a sample contains outliers or when the sample can logically be divided into several subsamples, each of which can be analyzed separately. Testing the difference of subsample averages may indicate whether the separation into subsamples can be taken as significant or not. We limit ourselves here to the situation where outliers exist in a data batch. Outliers differ significantly from all other values and can be readily identified by EDA plots. Outliers cause distortion of the $\bar{x}$ and s^2 estimates and may impair the subsequent statistical testing.

There are many different techniques, for example, cf. Ref. [5], for identifying outliers when a normal distribution of data can be assumed. One of the simplest and most efficient methods seems to be Hoaglin's modification of inner bounds B_L^* and B_U^*

$$B_L^* = \tilde{x}_{0.25} - K(\tilde{x}_{0.75} - \tilde{x}_{0.25}) \tag{4a}$$

and

$$B_U^* = \tilde{x}_{0.75} + K(\tilde{x}_{0.75} - \tilde{x}_{0.25}) \tag{4b}$$

where $\tilde{x}_{0.25}$ is the lower quartile, $\tilde{x}_{0.75}$ is the upper quartile and the value of parameter K is selected such that the probability $P(n, K)$ that no observation from a sample of size n will lie outside the modified inner bounds $[B_L^*, B_U^*]$ is sufficiently high, for example, $P(n, K) = 0.95$. For $P(n, K) = 0.95$ and $8 \le n \le 100$, Hoaglin [4] derived the following equation for the calculation of K:

$$K \approx 2.25 - (3.6/n). \tag{4c}$$

All elements lying outside the modified inner bounds $[B_L^*, B_U^*]$ are considered to be potential outliers. This test cannot be used for outliers detection for non-normal data distribution.

2.5. Data transformation

When the EDA indicates that the sample distribution strongly differs from a normal one, we are faced with the problem of how to analyze the biochemical or medical data. When examining data we must often find the proper transformation leading to symmetric data distribution, stabilizing the variance, or making the distribution closer to normal. Transformation for symmetry is carried out by a simple *power transformation*

$$y = g(x) = \left\{\begin{array}{ll} x^{\lambda} & \text{for parameter } \lambda > 0 \\ \ln x & \text{for parameter } \lambda = 0 \\ -x^{-\lambda} & \text{for parameter } \lambda < 0 \end{array}\right\} \quad (5)$$

which does not retain the scale, is not always continuous, and is suitable only for positive x. Optimal estimates of parameter $\hat{\lambda}$ are sought by minimizing the absolute values of particular characteristics of asymmetry. Transformation leading to approximate normality may be carried out by the family of Box–Cox transformations defined as

$$y = g(x) = \left\{\begin{array}{ll} (x^{\lambda} - 1)/\lambda & \text{for parameter } \lambda \neq 0 \\ \ln x & \text{for parameter } \lambda = 0 \end{array}\right\} \quad (6)$$

where x is a positive variable and the power λ is a real number. The Box–Cox transformation can be applied only on positive data. To extend this transformation means to make a substitution of x values by $(x - x_0)$ values which are always positive. Here x_0 is the threshold value $x_0 < x_{(1)}$. When the value 1 is covered by the confidence interval of estimated power λ, the data transformation is not efficient.

After an appropriate transformation of the original data $\{x\}$ has been found so that the transformed data give an approximately normal symmetrical distribution with constant variance, the statistical measures of location and spread for the transformed data $\{y\}$ are calculated. These include the sample arithmetic mean $\bar{y}$, the sample variance $s^2(y)$, and the confidence interval of the mean $\bar{y} \pm t_{1-\alpha/2}(n-1)s(y)/\sqrt{n}$. These estimates must then be recalculated for the original data $\{x\}$. The correct re-expression is based on the Taylor series expansion of the function $y = g(x)$ in the neighbourhood of the value $\bar{y}$. The re-expressed mean $\bar{x}_{\mathrm{R}}$ is then given

$$\bar{x}_{\mathrm{R}} \approx g^{-1}\left\{\bar{y} - \frac{1}{2}\frac{\mathrm{d}^2 g(x)}{\mathrm{d}x^2}\left(\frac{\mathrm{d}g(x)}{\mathrm{d}x}\right)^{-2} s^2(y)\right\} \quad (7)$$

For the variance it then holds

$$s^2(x_{\mathrm{R}}) \approx \left(\frac{\mathrm{d}g(x)}{\mathrm{d}x}\right)^{-2} s^2(y), \quad (8)$$

where individual derivatives are calculated at the point $x = \bar{x}_{\mathrm{R}}$. The $100(1-\alpha)\%$ confidence interval of the re-expressed mean for the original data is

$$I_{\mathrm{L}} \leq \mu \leq I_{\mathrm{U}} \quad (9)$$

where

$$I_{\mathrm{L}} = g^{-1}\left[\bar{y} + G - t_{1-\alpha/2}(n-1)\frac{s(y)}{\sqrt{n}}\right] \quad (10)$$

$$I_{\mathrm{U}} = g^{-1}\left[\bar{y} + G + t_{1-\alpha/2}(n-1)\frac{s(y)}{\sqrt{n}}\right] \quad (11)$$

and

$$G = -\frac{1}{2}\frac{\mathrm{d}^2 g(x)}{\mathrm{d}x^2}\left(\frac{\mathrm{d}g(x)}{\mathrm{d}x}\right)^{-2} s^2(y). \quad (12)$$

On the basis of the (known) actual transformation $y = g(x)$ and the estimates $\bar{y}$, $s^2(y)$, it is easy to calculate re-expressed estimates $\bar{x}_{\mathrm{R}}$ and $s^2(\bar{x}_{\mathrm{R}})$:

1. For a logarithmic transformation (when $\lambda = 0$) and $g(x) = \ln x$, the re-expressed mean and variance are calculated

$$\bar{x}_{\mathrm{R}} \approx \exp[\bar{y} + 0.5 s^2(y)] \quad (13)$$

and

$$s^2(\bar{x}_{\mathrm{R}}) \approx \bar{x}_{\mathrm{R}}^2(y) s^2(y). \quad (14)$$

2. For $\lambda \neq 0$ and the Box–Cox transformation, the re-expressed mean $\bar{x}_{\mathrm{R}}$ will be represented by one of the two roots of the quadratic equation

$$\bar{x}_{\mathrm{R},1,2} = [0.5(1 + \lambda\bar{y}) \pm 0.5\sqrt{1 + 2\lambda(\bar{y} + s^2(y)) + \lambda^2(\bar{y}^2 - 2s^2(y))}]^{1/\lambda} \quad (15)$$

which is close to the median $\tilde{x}_{0.5} = g^{-1}(\tilde{y}_{0.5})$. If $\bar{x}_R$ is known, the corresponding variance may be calculated from

$$s^2(x) = \bar{x}_R^{(-2\lambda+2)} s^2(y). \quad (16)$$

Note:

As an alternative to transformation, the maximal likelihood method may be used to find parameters of a selected distribution (e.g. log-normal, gamma, Cauchy), and the mean value calculated analytically. This method, however, needs the type of distribution to be known, is rather elaborate for practical use and sometimes the optimization may fail to find the correct parameter values.

Some other robust estimates of mean values based on influence functions, such as iteratively calculated M-estimates, will give similar values to $\bar{x}_R$, but the confidence intervals are symmetric and they assume a symmetric underlying distribution with several outliers.

3. Procedure

3.1. Procedure of extended exploratory data analysis

The extent of EDA and CDA of univariate biochemical data is best chosen according to experience from prior data analysis [8]. We consider here two common situations: (a) the treatment of routine data; and (b) the treatment of new data when no preliminary information is available.

3.1.1. The analysis of routine data

With routine biochemical data, some knowledge of the sample distribution is assumed—it is usually normal, and the data elements are homogeneous and independent. Tests for examining all assumptions about data should include: (i) a test for the independence of sample elements; (ii) a test for normality; (iii) a test for the homogeneity of the sample. Graphical EDA techniques such as the rankit plot and quantile–box plot are often used. When no preliminary information about the data is available, the full range of EDA plots should be followed by determination and construction of the sample distribution. Where no suitable distribution is found, the power transformation of data is recommended [6].

3.1.2. The analysis of new data

Analyzing a new data batch, there are several cases that require different strategies for the EDA procedures:

3.1.2.1. Data Case I. No independence of sample elements: when the sample elements are not proved to be independent, a danger of systematically biased and over-evaluated estimates arises. Therefore, a new logical analysis of the experimental equipment and data measurement procedures is necessary: after an improvement in the experimental strategy, the new data should be examined again. Alternatively, the data should be treated as a time series and a time series model should be found and explained.

3.1.2.2. Data Case II. No normality of sample distribution: the actual sample distribution is not normal in nature, or outliers are present in the data. When the distribution is not normal, the deviation can be in the lengths of tail (kurtosis) or in non-zero skewness. When tails differ in length, robust estimates may be used; for skewed distributions, a power transformation should always be used. When a power transformation is successful and the optimal value of the power $\hat{\lambda}$ is found, the estimates of the measures of location and spread can be calculated and re-expressed in the measure of the original variables.

3.1.2.3. Data Case III. Sample is not homogeneous, outliers: it should first be considered whether the distribution is skewed or not, because some points appearing to be outliers in a symmetrical (normal) distribution would be accepted in a skewed distribution. When some points are suspected outliers there are two alternatives: (a) to exclude the outliers from the data batch, which for a small sample size may lead to a loss of valuable information; or (b) to apply robust methods. In some biochemical or clinical data batches no outliers may be excluded because of the danger of losing valuable information. In such cases the experimenter should be consulted about the suspect points from a physical point of view, in order to consider the possibility of gross errors. Outliers can completely distort descriptive statistics. Comparing the mean, median, mode, and trimmed mean if the outliers are only to one side of the mean, the median is a better measure of location. On the other hand, if the outliers are equally placed on each side of the center, the mean and median will be close together, but the standard deviation will be inflated. The interquantile range is the only measure of variation not greatly, if at all, affected by outliers. Outliers may also contaminate measures of skewness and kurtosis as well as confidence limits.

3.1.2.4. Data Case IV. The sample size is not sufficient, missing data: whenever data are missing, question need to be asked: (a) Is the absence due to incomplete data collection? if so, try to complete the data collection; (b) Is the absence due to non-response from a survey? If so, attempt to collect data from the non-responders; (c) Are the absence data due to a censoring of data beyond or below certain values? If so, some different statistical tools will be needed; (d) Is the pattern of absence random? If only a few data points are missing from a large data set and the pattern of absence is random, there is little to be concerned with; however, if the data set is small or moderate in size, any degree of absence could cause bias in interpretations.

Where missing values occur without positive answers to the above questions, there is little that can be done: the best solution is to carry out new experimental measurements. As a general rule, when the variance of the data is small, a relatively smaller size will be required for any given precision of estimate. When no extra experiments can be carried out, the technique for small sample sizes should be applied. This is convenient for routine data analysis, but for new data exploratory data analysis should be used first, so that any statistical peculiarities of the sample may be determined.

3.2. Software used

For the creation of EDA diagnostic graphs and the computation of the quantile based characteristics of sample distribution, the EDA algorithm in S-Plus was written. This enables a test of sample independence based on the autocorrelation coefficient, a test of normality and a test of homogeneity based on the normality assumption, too. The original EDA algorithm is available on the internet at http://www.trilobyte.cz/EDA.

4. Results and discussion

Many statistical programs offer a list of various descriptive statistics of location and spread, but rarely help the user to choose one adequate measure for an actual sample batch. EDA and an examination of sample assumptions in CDA will find an answer to this question. The case study runs on typical biochemical sample data and illustrates the rigorous procedure of statistical data treatment with EDA and CDA, i.e. the exact assesment of the mean-value and the variance in 17-hydroxypregnenolone. Lower levels of free 5-ene steroids in umbilical blood and elevated levels of 5-ene steroid sulfates indicate a congenital sex-specific placental sulfatase insufficiency [9]. Delayed onset of labor, frequently linked with the necessity of intervention [10] together with relatively low birth weights is a common symptom of the disease. The defect of recessive X-linked type, also called the 'dry skin' disease, may have phenotypic consequences in later postnatal life [11]. The incidence of this disorder appears to be approximately one per 2000 male births [12]. The exact assessment of the mean-value and the variance of steroid levels in controls is necessary for the correct assessment of samples from patients. Lower levels of 5-ene steroid sulfates are common in pregnancies complicated by intrauterine fetal growth retardation (IUGR) [13]. The levels of pregnenolone, 17-hydroxypregnenolone in the umbilical blood of newborns, were evaluated. The evaluation of levels of 17-hydroxypregnenolone was chosen as an example of correct data analysis. Statistical assumptions should be tested on the

Table 1
The concentration of 17-hydroxypregnenolone in the umbilical blood of newborns (nmol/l) for sample size $n = 100$

17.7	19.6	35.1	29.1	23.7	25.7	35.6	21.3	30.5	28.0	25.6	155.6	18.0	21.6
16.8	34.6	21.8	37.2	31.8	33.3	25.6	25.8	26.7	22.9	19.2	23.1	26.4	15.3
37.8	36.7	35.4	11.8	5.0	36.2	19.2	16.2	21.9	30.8	34.2	20.9	24.2	18.7
32.1	15.5	59.2	13.7	17.9	16.5	20.8	20.5	19.1	32.4	38.6	29.8	24.2	22.2
16.7	27.2	21.8	24.8	7.9	6.0	8.1	4.0	21.5	21.0	17.4	16.5	19.8	19.7
15.7	21.7	23.7	23.3	23.3	41.1	29.1	21.8	23.0	19.5	25.3	41.6	31.3	24.0
44.3	22.3	35.9	29.8	16.5	33.9	24.3	20.3	20.9	17.8	28.0	33.6	22.9	22.4
27.1	23.6												

group of umbilical blood from newborns using some plots of extended exploratory data analysis and the statistical tests of basic assumptions. The estimate of a mean value in 17-hydroxypregnenolone was enumerated (Table 1).

(1) *Descriptive statistics:* NCSS2000 [14] software calculates a survey of descriptive statistics of location and spread for an actual sample size $n = 100$ (an elucidation of statistics cf. Ref. [6]). However, for the user it is rather difficult to select the single, the most convenient measure of the exact mean-value: the arithmetic mean $\bar{x} = 25.8$ nmol/l, the median $\hat{x}_{0.5} = 23.2$ nmol/l, the geometric mean $\bar{x}_g = 23.1$ nmol/l, the harmonic mean $\bar{x}_h = 20.4$ nmol/l, the mode $\hat{x}_M = 21.8$ nmol/l, and following trimmed means $\bar{x}(5\%) = 24.5$ nmol/l with $s(5\%) = 6.5$ nmol/l and $n(5\%) = 90$, $\bar{x}(10\%) = 24.4$ nmol/l with $s(10\%) = 5.5$ nmol/l and $n(10\%) = 80$, $\bar{x}(25\%) = 23.7$ nmol/l with $s(25\%) = 2.8$ nmol/l and $n(25\%) = 50$, $\bar{x}(45\%) = 23.2$ nmol/l with $s(45\%) = 0.4$ nmol/l and $n(45\%) = 10$; a survey of measures of spread: variance $s^2 = 248.4$, S.D. $s = 15.8$ nmol/l, unbiased S.D. $s = 15.8$ nmol/l, interquantile range $R_F = 11.1$ nmol/l; and a survey of measures of shape: skewness $\hat{g}_1 = 5.70$, kurtosis $\hat{g}_2 = 47.37$.

(2) *Exploratory data analysis* were used for the graphical visualization of 17-hydroxypregnenolone data: the quantile plot (Fig. 1) shows a small systematic deviation from a normal distribution and two outliers are detected. The classic curve for a normal distribution differs from the empirical, robust one. Both dot diagrams (Fig. 2) and the box-and-whisker plot (Fig. 3) indicate a nearly asymmetric distribution with two outliers. The halfsum plot (also called the midsum plot) (Fig. 4) and the symmetry plot (Fig. 5) indicate an asymmetric distribution with many points outsid the tested confidence bound.

(3) *Determination of sample distribution:* th sample distribution represented by a symmetr skewness and kurtosis is examined by four plot the histogram (Fig. 6) exhibits an asymmetri distribution of the sample analyzed. The kern density estimator of the probability density func

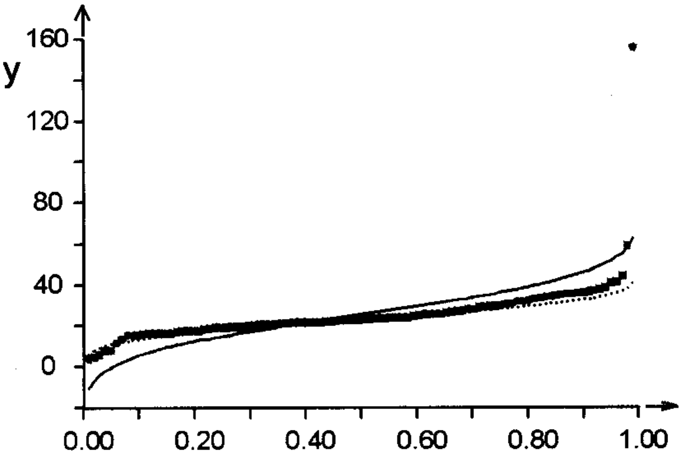

Fig. 1. The quantile plot for 17-hydroxypregnenolone data.

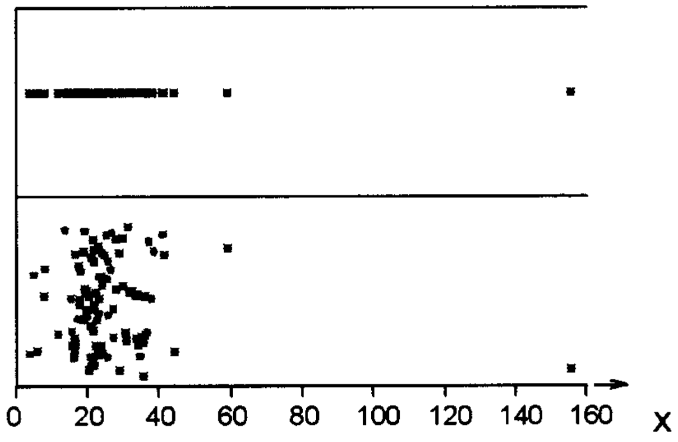

Fig. 2. The dot and jitter dot diagram for 17-hydroxypreg nenolone data.

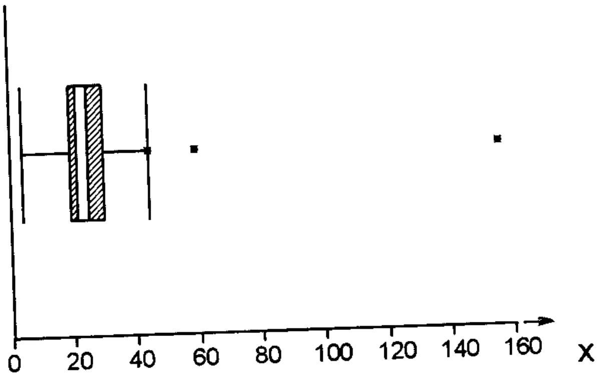

Fig. 3. The box-and-whisker plot for 17-hydroxypregnenolone data.

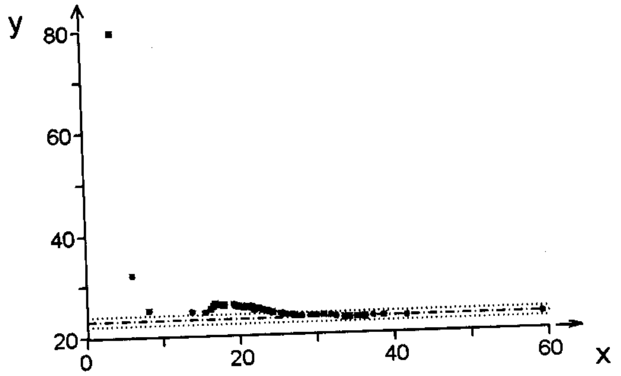

Fig. 4. The halfsum plot for 17-hydroxypregnenolone data.

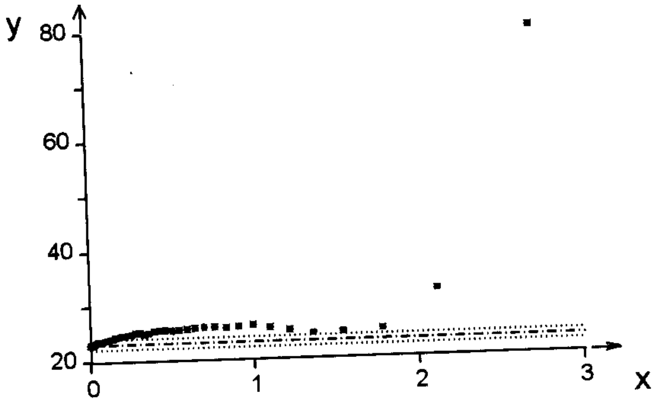

Fig. 5. The symmetry plot for 17-hydroxypregnenolone data.

tion (Fig. 7) does not prove a normal distribution as both curves, the theoretical of a normal distribution and the sample, significantly differ. The rankit plot, i.e. the quantile–quantile plot for normal distribution (Fig. 8), does not prove a normal distribution because not all the points are located on a straight line. In regression analysis of the quantile–quantile plot the various distributions with the sample one are compared, and the highest value of correlation coefficient $r = 0.89897$ is for a log-normal distribution. The quantile–box plot (Fig. 9) indicates many outliers outside the sedecile box.

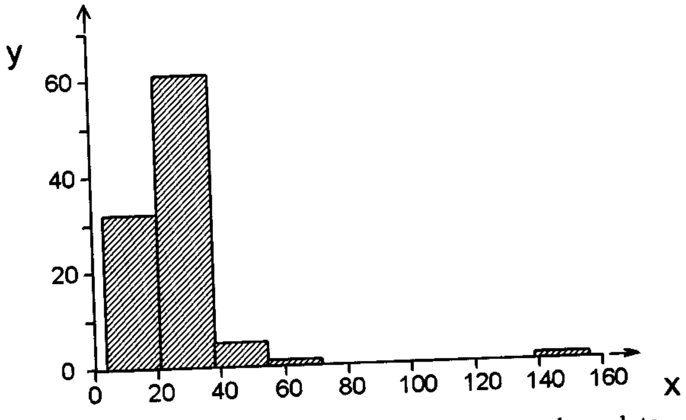

Fig. 6. The histogram for 17-hydroxypregnenolone data.

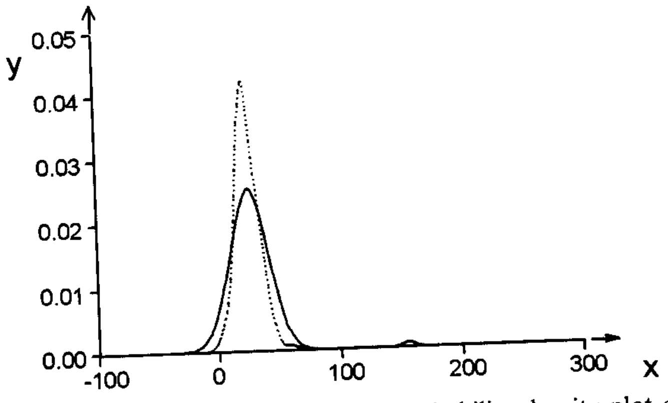

Fig. 7. The kernel estimator of the probability density plot of 17-hydroxypregnenolone data.

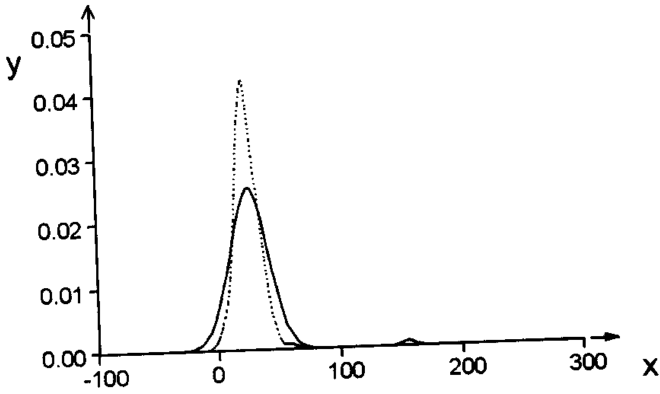

Fig. 8. The quantile–quantile plot for 17-hydroxypregnenolone data.

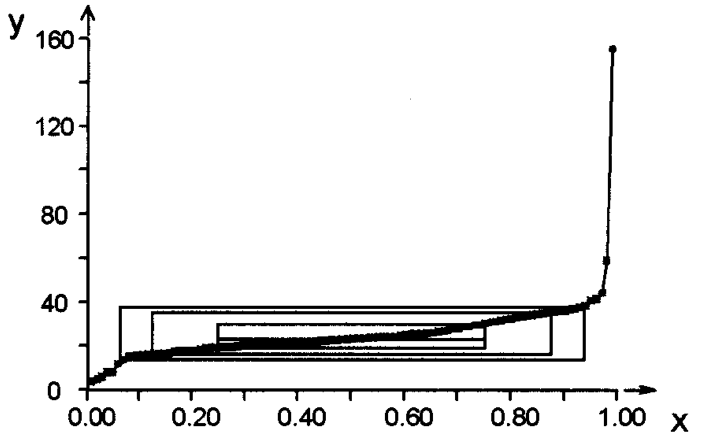

Fig. 9. The quantile–box plot for 17-hydroxypregnenolone data.

Tails length $T_E = 0.577$ is close to the tabular value for a normal distribution $T_E = 0.534$ and $T_D = 0.809$ is also close to the tabular value for a normal distribution $T_D = 0.822$ (Table 2). However, two outliers at higher values cause a log-normal distribution. The both point estimate of a skewness is 5.70 (for a normal distribution it should be zero) and of a kurtosis 47.37 (for a normal distribution it should be 3), indicating that the sample distribution has a strongly asymmetric, skewed and sharp shape.

(4) *Basic assumptions about the sample:* applying an analysis of basic assumptions about data the following conclusions were drawn:

(a) *Examination for independence of sample elements:* a test of sample elements independence leads to the test statistic $t_{17} = 0.05$ being lower than $t_{0.975}(100) = 1.984$ and therefore independence of sample elements is accepted.

(b) *Examination for normality of sample distribution:* a combined sample skewness and kurtosis test leads to the test statistic $C_1 = 10\,121$ being higher than $\chi^2(0.95, 2) = 5.992$ and therefore normality is rejected.

(c) *Examination of sample homogeneity:* there are two observations outside the interval of both of Hoaglin's inner bounds [$B_L^* = -4.6$ nmol/l; $B_U^* = 54.1$ nmol/l], i.e. $x_{(12)} = 155.6$ nmol/l, $x_{(45)} = 59.2$ nmol/l may be indicated as outliers and sample homogeneity is rejected. The measures of location, spread and distribution shape for sample *data without the two outliers* are $\bar{x} = 24.1$ nmol/l, $s(\bar{x}) = 8.1$ nmol/l, $\hat{g}_1(x) = 0.08$ and $\hat{g}_2(x) = 3.12$ leading to a the conclusion that the sample 17-hydroxypregnenolone without the two outliers exhibits a normal distribution.

(5) *Conclusions about the sample:* Table 3 provides a criticism of various measures of location, spread and distribution shape, [6]. All EDA display techniques prove that the sample distribution comes from a population with a log-normal distribution and that the sample contains two outliers. The classical measures of location and spread for the original sample data reaching values of mean $\bar{x} = 25.8$ nmol/l and standard deviation $s(x) = 15.8$ nmol/l are out of statistical significance and can not be used. For biochemical and clinical data a general rule is valid: *No outliers can be excluded from the sample batch because of the danger of losing valuable information;* therefore, no trimmed means can be used. For the best point estimate of the measure of location, using the Box–Cox and power transformation leading to the same estimates [6], the re-transformed mean $\bar{x}_R = 23.5$ nmol/l with the confidence interval $L_L = 21.4$ nmol/l and $L_U = 25.7$ nmol/l, and for the measure of spread the S.D. $s = 10.8$ nmol/l may be used. Besides the corrected mean by the data transformation the robust estimate median M with the

Table 2
The quantile measures of the location, spread and shape of 17-hydroxypregnenolone in the umbilical blood of newborns (nmol/l)

Quantile	P	Lower quantile Q_L	Upper quantile Q_U	Range R_Q	Halfsum Z_Q	Skewness S_Q	Tails length T_Q	Pseudo-sigma G_Q
Median	0.5	23.20	23.20	–				
Quartile	0.25	19.42	29.98	10.55	24.70	1.46	0.000	7.83
Octile	0.125	16.50	35.29	18.79	25.89	0.80	0.577	8.17
Sedecile	0.0625	14.00	37.69	23.69	25.84	0.65	0.809	7.74

Table 3
A survey of the critically commented measures of location, spread and shape of 17-hydroxypregnenolone

Measure	Point estimate	Lower limit	Upper limit	Criticism of measure used
Arithmetic mean	25.8	22.7	28.9	Not possible for asymmetric distribution
Median	23.2	21.4	25	The best estimate for asymmetric distribution
Geometric mean	23.1			The best estimate for log-normal distribution
Harmonic mean	20.4			Not possible for log-normal distribution
Trimmed mean (5%)	24.5	22.9	26.2	Unusable for biochemical data
Trimmed mean (10%)	24.5	22.9	26.2	Unusable for biochemical data
Trimmed mean (40%)	23.9	22.1	25.7	Unusable for biochemical data
M-estimate	24.1	22.3	25.8	A good estimate for asymmetric distribution
Hoog *M*-estimate	23.3	21.8	24.7	A good estimate for asymmetric distribution
Re-expressed mean after Box–Cox transformation	23.5	21.4	25.7	The best estimate for asymmetric distribution
Re-expressed mean after power transformation	23.5	21.4	25.7	The best estimate for asymmetric distribution
S.D.	15.8			Not possible for asymmetric distribution
Skewness	5.70			Indicates an asymmetrical, skewed distribution
Kurtosis	47.37			Indicates a sharp peak of distribution shape

confidence interval $M \pm 1.57R_L/\sqrt{n}$ also may be used, 23.2 ± 1.5 nmol/l, i.e. $L_L = 21.7$ nmol/l and $L_U = 24.7$ nmol/l.

5. Conclusions

The classic approach to instrumental data analysis in practice is based on some strong assumptions about the statistical nature of the data, such as the independence of sample elements, sample normality, sample homogeneity, and minimal sample size. To obtain undistorted and accurate results from univariate biochemical data, exploratory data analysis (EDA) should be applied to reveal typical features and patterns. Often, biochemical data are less than ideal and do not fulfill all these assumptions. For biochemical and clinical data no outliers can be excluded from the sample because of the danger of losing valuable information: robust statistics (median or *M*-estimates) or data transformation methods are recommended. The original data are then transformed to improve the symmetry of data distribution. Statistical measures of transformed data are re-transformed to obtain these unbiased and rigorous measures for the original data.

Acknowledgements

The financial support of the Grant Agency of the Czech Republic (Grant No 303/00/1559) and of the Ministry of Education (Grant No MSM253100002) is gratefully acknowledged.

References

[1] J.W. Tukey, Exploratory Data Analysis, Addison Wesley, Reading, MA, 1977.

[2] J. Chambers, W. Cleveland, W. Kleiner, P. Tukey, Graphical Methods for Data Analysis, Duxbury Press, Boston, 1983.

[3] D.C. Hoaglin, F. Mosteler, J.W. Tukey, Exploring Data Tables, Trends and Shapes, Wiley, New York, 1985.

[4] D.C. Hoaglin, F. Mosteler, J.W. Tukey (Eds.), Understanding Robust and Exploratory Data Analysis. Wiley, New York, 1983.

[5] K. Stoodley, Applied and Computational Statistics, Ellis Horwood, Chichester, 1984.

[6] M. Meloun, J. Militký, M. Forina, Chemometrics for Analytical Chemistry, Part 1. PC-aided Statistical Data Analysis, Ellis Horwood, Chichester, 1992.

[7] M. Meloun, J. Militký, M. Forina, Chemometrics for Analytical Chemistry, Part 2. PC-aided Regression and Related Methods, Ellis Horwood, Chichester, 1994.

[8] ADSTAT statistical package, TriloByte Statistical Software Ltd., Pardubice, 1998.

[9] K. Hirato, T. Yanaihara, Serum steroid hormone levels in neonates born from the mother with placental sulfatase deficiency, Endocrinol. Jpn. 37 (1999) 731.

[10] T. Rabe, R. Hosch, B. Runnebaum, Sulfatase deficiency in the human placenta: clinical findings, Biol. Res. Pregnancy Perinatol. 4 (1983) 95.

[11] L.J. Shapiro, L. Cousins, A.L. Fluharty, R.L. Stevens, H. Kihara, Steroid sulfatase deficiency, Pediatr. Res. 11 (1977) 894.

[12] G. Lykkesfeldt, M.D. Nielsen, A.E. Lykkesfeldt, Placental steroid sulfatase deficiency: biochemical diagnosis and clinical review, Obstet. Gynecol. 64 (1984) 49.

[13] C.R. Parker Jr, E.S. Buchina, T.K. Barefoot, Abnormal adrenal steroidogenesis in growth-retarded newborn infants, Pediatr. Res. 35 (1994) 633.

[14] NCSS Statistical Software, 329 North 1000 East, Kaysville, Utah 84037. Email: sales@mcss.com.